USING START/END TIMINGS OF SPECTRAL TRANSITIONS BETWEEN PHONEMES IN CONCATENATIVE SPEECH SYNTHESIS

Toshio Hirai[†‡] *Seiichi Tenpaku*[†] *Kiyohiro Shikano*[‡]

† Arcadia Inc., 562–0003–3–1–15 Japan
‡ Nara Institute of Science and Technology, 630–0101–8916–5 Japan
thirai@arcadia.co.jp

ABSTRACT

The definition of "phoneme boundary timing" in a speech corpus affects the quality of concatenative speech synthesis systems. For example, if the selected speech unit is not appropriately match to the speech unit of the required phoneme environment, the quality may be degraded. In this paper, a dynamic segment boundary definition is proposed. In the definition, the concatenation point is chosen from the start or end timings of spectral transition depending on the phoneme environment at the boundaries. For a listening test to compare the naturalness of conventional/proposed methods, 100 Japanese place names were selected randomly and synthesized. The ratio of naturalness was 1 to 3.3 (conventional v.s. proposed) by four subjects.

1. INTRODUCTION

A speech synthesis system is one means by which humans can communicate smoothly with machines. In the construction of such a system, two steps are carried out: (1) Record speech data which includes speech units (phonemes, syllables, etc.) used at the synthesis stage; (2) Analyze the units into acoustical parameters and register them into a database. At the synthesis stage, required (but unrecorded) speech is generated by selecting and concatenating the suitable units with or without signal processing. Recently, speech concatenative systems [1, 2] are being widely developed, and speech units are concatenated as they are.

In this method, when appropriate units are selected, the quality of the synthesized speech is high enough. Additionally, since the speech is constructed from the original speech sounds, the individuality of the recorded speaker is easily reproduced. The appropriateness or inappropriateness of a speech unit is affected by the following three elements: (1) the size of the speech database, (2) the algorithm for searching for and selecting the appropriate speech unit, and (3) the features which are used at the search algorithm. Paraphrasing, if the size of the speech database is large enough, then the variation of speech units increases, and appropriate units can be found easily. If the algorithm and the features describing the speech unit when searching are suitable, it is possible to evaluate the unit's appropriateness accurately in the synthesized speech, and the quality of the speech improves. However, these conditions are not often satisfied. Especially, the size of the speech database tends to be inadequate, so speech synthesis often takes place using fewer speech units than required.

A typical example of the inappropriate selection of speech units is when the phoneme environments of the speech unit pair's connection ends do not agree. Kawai et al. proposed a method in which a list of the connection costs was employed based on the evaluation of qualitative deterioration of synthesized speech which were then entered into a unit search algorithm [3]. In the synthesis, the connection timing information between vowels ('V', the last phoneme of previous unit) and consonants ('C', the first phoneme of the successive unit) is determined rigidly whether the phoneme environments agree with each other or not. If the environments do not agree, the problem described below occurs.

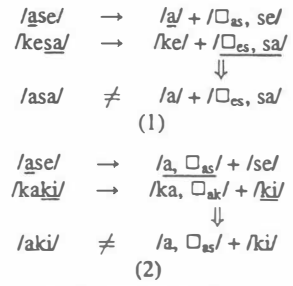

Fig. 1. Problems in conventional synthesis method. Unit boundary is defined as the start timing (case (1), /asa/ synthesis) or as the end timing (case (2), /aki/) of the transitions.

For example, in the case of a concatenative synthesis system based on mora units ("mora (*pl.* morae)" is a unit of CV in Japanese), if the speech /asa/ ("morning" in Japanese) is synthesized from the first mora of /ase/ ("sweat") and the second mora of /kesa/ ("this morning"), the connection is executed at the phoneme boundary between /a/ and /s/ of /asa/ and the boundary between /e/ and /s/ of /kesa/. If the phoneme boundary is defined as "the start timing of the fricative sound /s/ in the speech," then the part of the transition period (some pitches at the end of /kesa/'s /e/) is included in the synthesized speech. As the result, the voiced sound /e/ appears in the synthesized speech slightly, and the quality degrades. In order to solve this problem, if the definition of the phoneme boundary is changed to "the end timing when the voiced component disappears," the quality of the synthesized speech explained above would improve. However, in other synthesis cases, the quality degrades owing to the change of the boundary definition. For example, in the case of /aki/ ("autumn") synthesized from the first mora of /ase/ and the second mora of /kaki/ ("persimmon"), if we adopt the "improved" boundary definition, the fricative component of /ase/ is included in the synthesized speech, and again, the quality degrades. These situations are illustrated in **Fig. 1**. In the figure, "$\square_{ph_1\ ph_2}$" describes the "transition period from ph_1 to ph_2." For example, "$\square_{as}$" is the transition period from /a/ to /s/ in the case (1). It is clear that these problems are caused by the rigid use of boundary information. In order to solve the problems, we can introduce some kind of measure [4, 5] to

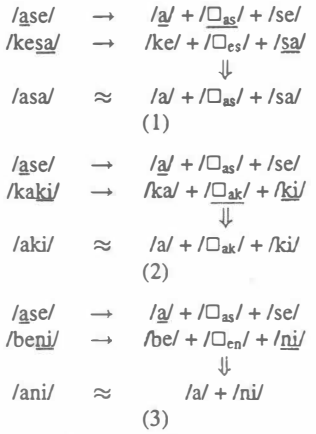

Fig. 2. Concatenation examples with proposed method. From the top: (1) the first phonemes of the successive units correspond with /s/, (2) the last phonemes of the previous units correspond with /a/, and (3) phoneme environments do not match (/a/ ≠ /e/ and /s/ ≠ /n/).

be applied for the optimization of the concatenation point dynamically. However, if the phoneme environment does not agree at the concatenation point (e.g. /ani/ from /ase/ and /beni/), the measure shows a large acoustical distance, and the optimization may not be executed properly.

2. FLEXIBLE USE OF START/END TIMINGS OF TRANSITION PERIOD IN SPEECH UNIT CONCATENATION

In this paper, in order to solve the problems described in the previous section, we propose a method in which the concatenation timing of the speech unit is determined not by the phoneme boundary timing but by the start/end timings of spectral transitions between phonemes depending on the phoneme environment of concatenated units. The transition period is defined as the point when the post phoneme first appears (start timing), and the end point when the previous phoneme disappears. The concatenation point is selected from these timings depending on the conformity of the end phonemes of adjacent speech units which are concatenated. The remaining part of this section gives some actual examples based on the proposed method.

In the case of /asa/ synthesis as described in the case (1) of Fig. 1, the first phonemes of the successive units are the same (/s/), so concatenation is expected at each end of transition. That is, from the speech /ase/, the part from the beginning of /a/ to the end of the transition between the phonemes /a/ and /s/ (/□as/) is clipped, and from the speech /kesa/, the part from the end of the transition /□es/ to the end of the speech is clipped, and they are connected to each other. Likewise, in the case of /aki/ synthesis from /ase/ and /kaki/ (case (2)), the last phonemes of the previous units are the same (/a/), and concatenation is executed at both beginning points of transition. Additionally, if the phoneme environments at the boundary do not correspond (like /ani/ from /ase/ and /beni/), the transition parts (□as and □en) are not used for the concatenation. The procedures are illustrated in **Fig. 2**. According to the proposed method, the transition period would be skipped at synthesis in case (3) in Fig. 2. Though it may cause the degradation of synthesized speech because of the change of phoneme duration, it is also expected that quality will improve due to the deletion of the unnecessary transition period.

Table 1. Records in the database from the place name "ao'mori"

current unit	phoneme	pos. and tone
ao	silB–ao+m	silB–1L,2H+3L
mo	o–mo+r	2H–3L+4L
ri	o–ri+silE	3L–4L+silE

3. GENERATION OF SPEECH DATABASE

3.1. Design of recording dataset and speech recording

Japanese place names were chosen as the synthesis target since such have been used often in template synthesis. The original data of the names were loaded from a home page of the Japanese Post Office [6], and accentual information was added using a semi-automatic method. The original number of names was about 120k, and the count of unique sequence patterns (phonetic and intonation) was 60,687. We call this the "synthesis target." Based on place names, we studied what kinds of speech units were needed for the task, and designed the recording dataset which was included in the units of the place names themselves and words whose range of length was the same as that of the place names (3–8 morae). We adopted the "greedy method" [7] for the design. Additionally, we added names which include rare speech units in order to be able to synthesize all the names in phonetic meaning. In this way, 4,004 names/words were selected for the recording. The recording list was constructed from the columns of the sequence number in the list, Chinese characters with "hiragana (the Japanese cursive syllabary)," and hiragana with accentual symbol. Speech data was collected using a female narrator. The recording list was separated into two parts (2,000 names/words each). The recording of each part took two days. The recording conditions were: soundproof chamber, monaural, digitally recorded 44.1 kHz sampling frequency, and 16-bit quantization.

3.2. Database production from the recorded speech

The speech data was down-sampled to 22.05 kHz and clipped into electronic files with 200 ms pauses attached at both ends of the utterances. The total size of the electronic files was 252 MB. Apart from the processing, the speech data was down-sampled to 16 kHz for automatic phonetic labeling by a continuous speech recognition system, Julius [8]. The labeling results were modified by hand as need arose, and the transition period information was added by hand according to in-house standards.

These factors were adopted in classifying speech units:

- the sequence of CV* (C + any length of V sequence) [9, 10] (hereafter "current unit"), pre/post-phoneme,
- the mora position of the current unit,
- and the accentual high/low information of the current unit and pre/post-mora ('H' and 'L' were used for the notation. In the case of type-0 accent which has no downfall of tone during the word, the higher tone part was notated as 'M' in order to distinguish it from 'H' of the high tone of non type-0 accent).

The reason why the CV* was chosen as the nucleus of the speech unit was to avoid unnatural concatenation midway through the vowel sequence. **Table 1** shows an example of records in the database for a place name sample "ao'mori." The apostrophe /'/ denotes the "accent nucleus" when the tone falls from H to L at the mark. The number and uppercase letter to the left of the "–" indicate the position and tone of the "previous" sound, and those to the right of the "+" indicate these of the "following" sound. The "silB" and "silE" denote the silence at the beginning/end of the

speech. The total number of current units was 16,668, and the number of unique units was 949. The total number of current units in the synthesis target was 45,219, and its 16.9% (7,644) were included in the recorded speech data. In each line of the record, the start/end timings of the current unit and the length of the unit in mora were included. If there was transition period at the ends of the unit boundaries, they were also notated.

Though the number of current unit types tends to be enormous under normal circumstances, if the target for synthesis is limited (as in this study), the amount is finite. H/L information and mora position were introduced as classification factors since the fundamental frequency of speech (F_0) is roughly the same in the group of units which have the same mora position and the same H/L [11]. Additionally, the high/low information of pre/post-mora was introduced for detecting the tendency of F_0 changes at the concatenation points.

4. SPEECH SYNTHESIS AND HEARING EXPERIMENTS

4.1. Synthesis procedure

Speech synthesis was executed by using the speech data described above. The factors described in 3.2 were adopted as criterion for unit selection. Additionally, the length of the synthesized name (which includes the current unit) was also adopted as a factor since the duration of the unit is affected by the length; a short word is usually uttered slowly, so the phoneme duration is long compared to a longer word. Though Mizusawa et al. proposed to adopt the length itself for the selection of unit as a factor [12], the number of factor classes would become massive and each class size would be small, so we chose to classify the lengths into a few categories.

In order to decide the number of categories, the relationship between the length of current unit and the length of word was investigated in two thousand words of the recorded data. **Figure 3** shows the results. The x-axes are the length of the word (which includes the current unit) in mora, and the y-axes are the duration of the unit in ms. The three plots describe 1 mora of the current unit, 2 morae, and more than 2 morae, respectively. R [13] is used for the analysis. According to the analysis, it is revealed that the group is split into two classes in each plot: (1) less than or equal to "the length of current unit + 2" morae, and (2) others. In the case where the current unit is 2 morae (upper right of Fig. 3), one class consists of "less than or equal to (2+2=) 4 morae word length" with the duration being 390 ms, and another class consists of "more than 4 morae word length" with the duration 310 ms, approximately. The classification was adopted in the speech synthesis system as a selection factor.

The speech signal was not touched except in the morphing processing (the pitch structure is changed gradually from the previous unit to the succeeding one) at the concatenated point in order to suppress the clicking noise caused by waveform concatenation.

4.2. Perceptual experiments

One hundred stimuli pairs were selected from the synthesis target whose units were the same but whose concatenation points were different between proposed and conventional methods. The boundary definition of the current unit in the conventional method was fixed at the end point of transition. Subjects played the sounds arbitrarily by pressing corresponding buttons on a HTML based GUI. The order of the methods (proposed or conventional) in a pair was arranged randomly. The subject was asked to judge which synthesized speech was natural. If the timing difference between the methods was not large enough, the difference of synthesized speech samples would be negligible, so the subjects were allowed

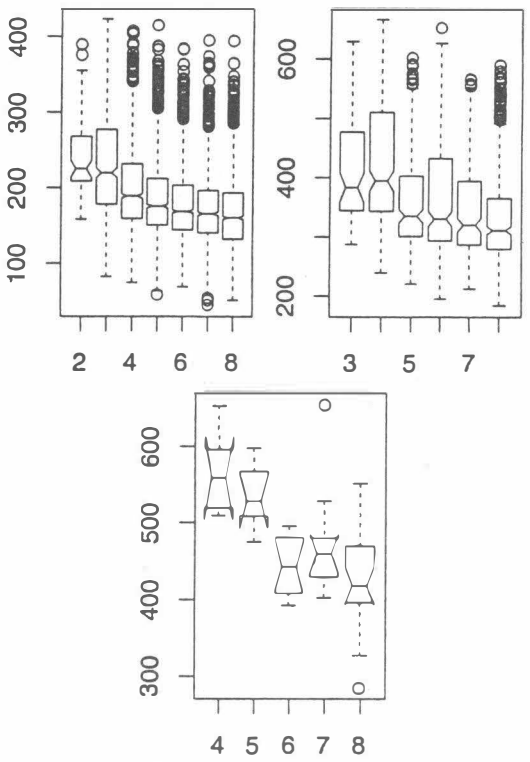

Fig. 3. The relationship between the current unit length and the word length.

Upper left: current unit is 1 mora, upper right: 2 morae, lower: more than 2 morae. The x-axes are the length of the word (which includes the current unit in mora), and the y-axes are the duration of the unit in ms.

Table 2. Preference scores of proposed/conventional methods

method	mean	S.D.
proposed	63.3	8.22
conventional	19.3	4.65

to judge the samples as having "equal naturalness." Two male and two female subjects, who were used to listening to synthesized speech, were used. The subjects listened with headphones under the noisy conditions of a business office. They were allowed to listen to the speech samples as many times as they wanted.

5. RESULTS

The means and standard deviations (S.D.) of preference scores are shown in **Table 2** as percentages. The ratio of naturalness was 1 to 3.3 (conventional v.s. proposed). In the experiments, neutral judgment for the naturalness was allowed, so the sum of the means does not total 100%. There were 39 cases in which all the subjects judged that the speech generated by the proposed method was more natural. Two cases received opposite judgment (all the subjects preferred the sound of the conventional method in these two cases). This is discussed in the next section.

6. DISCUSSION

The t-test was adopted to test if the means are significantly different. The t value at a degree of freedom of 3 ($= 4 - 1$) and a level of significance of 5% ($t_{3,0.025}$) is 3.18. The sum of the $t_{3,0.025} \times$ S.D.

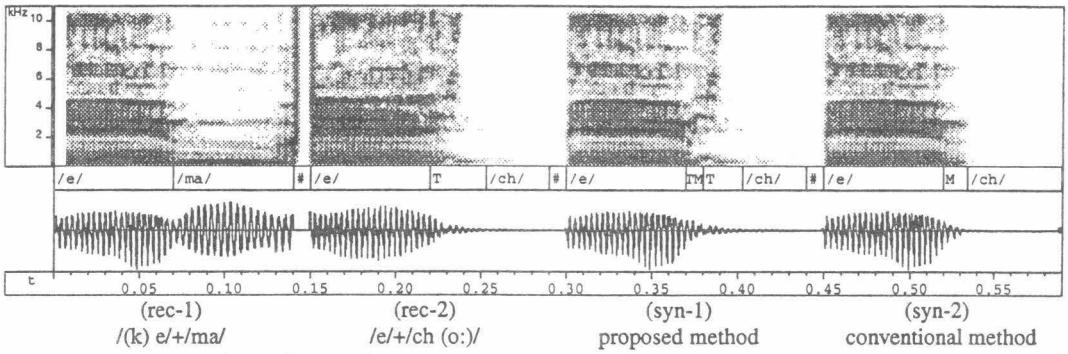

Fig. 4. Patched waveform of synthesized speeches /ech/ of "ShiNdeNbuke'cho:" and original speech samples.

of each method ($(8.22 + 4.65) \times 3.18 = 40.93$) was not enough to fill the difference between the means ($63.3 - 19.3 = 44.00$). Therefore, the difference was significant at the level; that is, the proposed method was significantly superior to the conventional one.

One of the two cases in which all the subjects preferred the sound of the conventional method was investigated. **Figure 4** is the patched waveform of the case with spectrogram, label, and time scale in seconds. The left two waveform fragments are part of the original (recorded) speech samples (/(k) ema/ (rec-1) and /ech (o:)/ (rec-2), ':' describes that the previous phoneme is a long vowel), and the right two fragments are the synthesized speech /ech/ in "ShiNdeNbuke'cho:" generated by proposed (syn-1) and conventional (syn-2) methods from two recorded speech samples rec-1 and rec-2. Wavesurfer [14] was used for the plotting. In the figure, label 'T' denotes the transition period and 'TM' and 'M' denote the morphed part ('TM' is in a transition period). Pauses ('#') are inserted between the fragments. Since the following phoneme of the selected unit /ke/ (/m/) does not correspond to the target phoneme (/ch/), and the preceding phoneme of the selected unit /cho:/ (/e/) corresponds to the target phoneme, the transition period between /e/ and /ch/ (the 'T' part of rec-2) was used in the proposed method. Comparing the spectrograms of synthesized speech samples syn-1 and syn-2, it can be concluded that the unnatural appearance of the high frequency component (6kHz–10kHz at 0.38s) might cause the judgment in the experiments. Even though the component in /ke/ in rec-1 fades away, the component in the transition period before /ch/ does not fade. This phenomenon suggests that the inconsistent labeling affects the quality of synthesized speech sensitively. Computational alignment may be effective for dealing with such problems.

7. SUMMARY

In this paper, a new method of label information usage for concatenative speech synthesis has been proposed. In this method, the start end timings of spectral transition between phonemes are used flexibly depending on the phoneme environment at the ends of selected speech units in concatenation. Hearing experiments were carried out using four subjects to evaluate the naturalness of the synthesized speech compared to the speech produced by the conventional method. The mean rate of the preference score was 63.3% for the proposed method and 19.3% for the conventional method. The ratio of naturalness was 1 to 3.3 (conventional v.s. proposed), and this difference is significant. We plan to improve the method by studying the mechanism further regarding why the conventional method was preferred in some samples. Additionally, it is important to establish a procedure for consistent labeling.

8. REFERENCES

[1] W.N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis*, J. P. H. Van Santen, J. Olive, J. Hirshberg, and R. Sproat, Eds., pp. 279–292. Springer, N.Y., 1997.

[2] A. Syrdal, C. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, and K. Lee, "Corpus-based techniques in the AT&T NEXTGEN synthesis system," in *Proc. ICSLP2000*, SS–05–02, Oct. 2000.

[3] H. Kawai, M. Tsuzaki, T. Masuda, and H. Iwasawa, "Perceptual evaluation of naturalness degradation due to substitution of phonetic environment for concatenative speech synthesis," Tech. Rep. SP2001–22, IEICE Jpn., May 2001, In Japanese.

[4] J. Wouters and M. W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, 1998, pp. 2747–2750.

[5] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, May 2001.

[6] "Zip code list," http://www.post.yusei.go.jp/newnumber/lzh/s/ken_all.lzh (as of June 7, 2002, In Japanese).

[7] V. Chvatal, "A greedy heuristic for the set-covering problem," *Mathematics of Operations Research*, vol. 4, no. 3, pp. 233–235, Aug. 1979.

[8] "Continuous Speech Recognition Consortium," http://www.lang.astem.or.jp/CSRC (as of June 7, 2002, In Japanese).

[9] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking account of prosody," in *Proc. ICSLP2000*, SS–05–04, Oct. 2000.

[10] K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, "A text-to-speech system using multi-form unit," in *Rec. Spring Meeting*. Acoust. Soc. Jpn., Sep.–Oct. 1999, pp. 211–212, In Japanese.

[11] H. Fujisaki and S. Nagashima, "A model for the synthesis of pitch contours of connected speech," Ann. Rep. Vol. 28, No. 53, Eng. Res. Inst., Univ. of Tokyo, 1969.

[12] N. Mizusawa, J. Murakami, and M. Higashida, "Simple word synthesis by concatenative syllabic components based on positional features with mora length," Tech. Rep. SP99–2, IEICE Jpn., May 1999, In Japanese.

[13] "The R Project for Statistical Computing," http://www.r-project.org/(as of June 7, 2002).

[14] "Wavesurfer," http://www.speech.kth.se/wavesurfer/(as of June 7, 2002).